



Moolan-Feroze, O., Karachalios, K., Nikolaidis, D., & Calway, A. (2019). Improving drone localisation around wind turbines using monocular model-based tracking. In *2019 IEEE International Conference on Robotics and Automation (ICRA 2019)* (pp. 7713 - 7719). (l: IEEE International Conference on Robotics and Automation). Institute of Electrical and Electronics Engineers (IEEE). <https://doi.org/10.1109/ICRA.2019.8794156>

Peer reviewed version

Link to published version (if available):
[10.1109/ICRA.2019.8794156](https://doi.org/10.1109/ICRA.2019.8794156)

[Link to publication record in Explore Bristol Research](#)
PDF-document

This is the author accepted manuscript (AAM). The final published version (version of record) is available online via IEEE at <https://ieeexplore.ieee.org/document/8794156>. Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research

General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available:
<http://www.bristol.ac.uk/pure/about/ebr-terms>

Improving drone localisation around wind turbines using monocular model-based tracking

Oliver Moolan-Feroze¹, Konstantinos Karachalios², Dimitrios N. Nikolaidis², and Andrew Calway¹

Abstract—We present a novel method of integrating image-based measurements into a drone navigation system for the automated inspection of wind turbines. We take a model-based tracking approach, where a 3D skeleton representation of the turbine is matched to the image data. Matching is based on comparing the projection of the representation to that inferred from images using a convolutional neural network. This enables us to find image correspondences using a generic turbine model that can be applied to a wide range of turbine shapes and sizes. To estimate 3D pose of the drone, we fuse the network output with GPS and IMU measurements using a pose graph optimiser. Results illustrate that the use of the image measurements significantly improves the accuracy of the localisation over that obtained using GPS and IMU alone.

I. INTRODUCTION

Due to harsh weather conditions, wind turbines can incur a wide range of structural damage [1], which can severely impact their power generation abilities [2]. To address this, regular inspections are needed. Current best practice in visual inspection is the use of ground-based cameras with telephoto lenses, or manual inspection using climbing equipment. Both methods incur considerable cost in both the inspection itself, and the turbine down time. Manual inspections can also lead to inconsistencies in the data gathering, which can be compounded over multiple visits.

To address the above, inspecting using unmanned autonomous vehicles (UAVs) – or drones – is being considered [3]. Autonomous inspections have the potential to save time and cost and give more consistent inspection data. A key element for successful inspections is the ability to accurately determine drone location with respect to the wind turbine. Accurate localisations increase the consistencies of inspections, allow the drone to get closer to the turbine, and are useful in image post-processing. Although global positioning systems (GPS) and inertial measurement units (IMUs) can provide relatively good tracking, improved performance can be obtained by inclusion of image-based sensor readings [4].

We present a novel system for integrating image-based measurements with GPS/IMU readings, which gives improved localisation of the drone. To do this we take a model-based tracking approach. An internal 3D skeleton representation of the wind turbine is matched to that inferred from

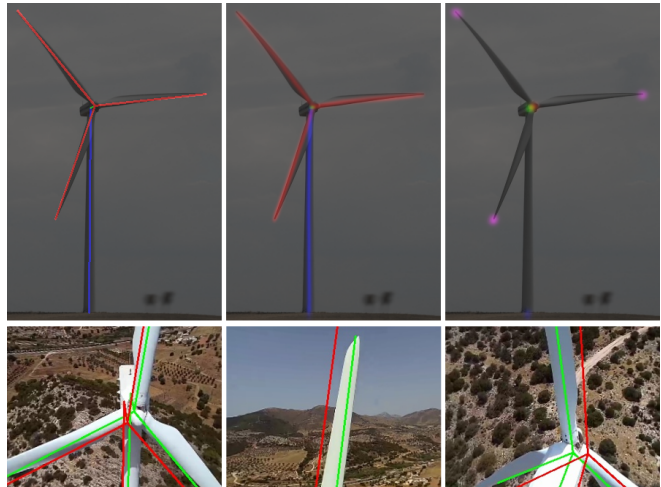


Fig. 1. Top Left) Wind turbine with the lines from \mathcal{L}^c overlaid. Top Middle) Line output from the CNN. Top Right) Point output from the CNN. Bottom) Example of the system applied to an inspection flight. Red lines are the reprojections of the model before optimisation. Green lines are the reprojections after optimisation.

image data using a convolutional neural network (CNN). The difference is minimised using a pose graph optimiser which is constrained by the GPS/IMU measurements.

There are two main contributions in this work. The first is a novel application of a CNN that is able to infer from image data the 2D projection of the 3D skeleton model. This enables us to easily find correspondences between the model and images. In addition, we incorporate prior information about the likely pose of the camera into the network to improve its prediction performance. The second contribution is the integration of the network output into a pose graph optimisation, using both point and line features.

In Section II, we explore some of the work related to the use of CNNs in localisation and tracking applications, as well as detailing the research into drone inspections of wind turbines. We present our method in Section III, and detail the turbine representation we have chosen, give a detailed description of the CNN, and describe how the CNN outputs are integrated with the pose optimiser. In Section IV we describe the evaluation of our method, using both real and simulated data. Finally, in Section V we give some conclusions and ideas for future work.

II. PREVIOUS WORK

Over recent years, deep learning has been applied to image-based camera pose estimation in a number of ways. These can be roughly separated into two groups: end-to-end approaches, and approaches that use deep learning as

¹Oliver Moolan-Feroze and Andrew Calway are with the Department of Computer Science, University of Bristol, 75 Woodland Road, Bristol, BS8 1UB, United Kingdom oliver.moolan-feroze@bristol.ac.uk, andrew.calway@bristol.ac.uk

²Konstantinos Karachalios and Dimitrios N. Nikolaidis are with Perceptual Robotics, 5 Hope Road, Bristol, UK, BS3 3NZ, United Kingdom kostas@perceptual-robotics.com, dimitris@perceptual-robotics.com

an intermediate or preprocessing step. The idea behind the end-to-end group is that given an input image, the 6DoF pose of the camera can be regressed directly by the network. The first attempt at this was PoseNet [5], where the authors designed a VGG [6] style CNN with a 3D translation regression block and 4D quaternion regression block as outputs. Through the use of transfer learning the authors trained the network for indoor and outdoor scenes using only a small number of pose-labelled images. This work was later expanded on with the inclusion of measures of uncertainty into the output [7] using Bayesian deep learning. Further additions were made in [8], including a novel geometrical loss function based on reprojection errors. Clark et.al. [9] take advantage of temporal smoothness in camera poses over neighbouring video frames through the use of Recurrent Neural Network (RNN) layers. A series of Long Short-term memory units are appended after a CNN block which are able to integrate features from previous time steps to aid in the regression process. This method is also used in [10], where the authors state the recurrent layers are able to learn the motion dynamics of the system over a period of time.

End-to-end approaches are beneficial in their simplicity, but there are a number of problems. The most obvious is the need for pose-labelled training data. This is at best expensive and time consuming to obtain and in some cases – such as in our application – impossible to gather. In addition, the performance of end-to-end methods is still lagging behind traditional geometric approaches.

One way of addressing these problems is to use deep learning as an intermediate or preprocessing step in a traditional geometric approach. In [11], the authors use a CNN to generate heatmaps of model feature points. The location of different features are represented by peaks in the heatmaps, and the pose of the camera is obtained through a minimisation process. The authors use a stacked hourglass [12] architecture which is able to integrate features from across the spatial extent of the input image. In [13], the authors propose a similar method of feature points extraction using CNNs. However, in this work they explicitly handle cases where the object is partially out-of-view. This is especially beneficial for tasks such as industrial inspection where due to close proximity, only incomplete views of the inspected object are visible.

There is not much literature related to the autonomous inspection of wind turbines using drones. Stokkeland et. al. [14] present a method to determine the position of the drone and the configuration of the wind turbine during an initial approach stage. Their method uses a Hough Transform to locate the different parts of the turbine, and then integrates this information through a Kalman Filter to track the position of the drone. One drawback of this work is that it only addresses the initial approach of the drone. The actual details of the inspection are not handled. The work in [15] propose the use of a LiDAR sensor to aid in navigation. It describes a 3D occupancy grid which is able to integrate multiple noisy sensor reading using a Bayesian update scheme. This grid then serves as a map for path planning and localisation. This

work has a number of important omissions however. First, it is not applied to real data, only performed in simulation. Second, it makes no attempt at localisation, focusing only on mapping and path planning.

III. METHOD

The process of localising the drone with respect to the wind turbine is split into two parts. During the first part (Section III-B), we obtain images of the turbine from the monocular camera on the front of the drone and pass them through a CNN to extract an estimate of the projection of the 3D skeleton representation. These estimates are then used to constrain a pose graph optimisation (Section III-C). Key to both these parts is how the wind turbine is represented by the system (Section III-A).

A. Turbine Representation

To enable a model-based tracking approach, an internal representation of the wind turbine is necessary. When estimating the pose of the camera, the representation is projected through the current estimate of the pose and the camera intrinsic matrix and then compared to the visual information contained within the image. The optimal pose is then the one that aligns the reprojection of the representation with the object in the image. We have chosen a very simple skeleton representation that is general enough such that it is able to fit to a wide range of different turbine shapes, sizes and configurations.

The representation is based on a set of 3D points, combined with a set of lines which connect them. The points \mathcal{P}^c lie at the base of the turbine tower, the top of the turbine tower, the centre of the blades, and the tips of each of the blades. These points were chosen by looking for the commonalities between different wind turbine shapes and sizes. The set of lines \mathcal{L}^c connect the bottom and the top of the tower, the top of the tower and the centre of the blades, and the centre of the blades with each of the blade tips. In total, this gives us a set of 6 points and 5 lines. An example of the representation can be seen in Figure 1. We assume that a good estimate of the size and shape of the turbine is known prior to applying the localisation process.

B. Model projection inference using CNNs

As described in Section III-A, the internal representation of the wind turbine is intentionally very general. However, as different wind turbines can present a wide range of visual information, it is difficult to find the accurate correspondences needed for localisation. Indeed, apart from the tips of the blades, none of the points in \mathcal{P}^c correspond to specific image features that would be common across all wind turbines. Furthermore, the lines in \mathcal{L}^c do not run along specific edges in the images, but rather through the centre of the different parts of the turbine. To address this, we make use of a CNN to process the input images into a form that can be easily matched to the projection of the skeleton model.

The network takes a multichannel image as an input and produces a multichannel image as an output using a

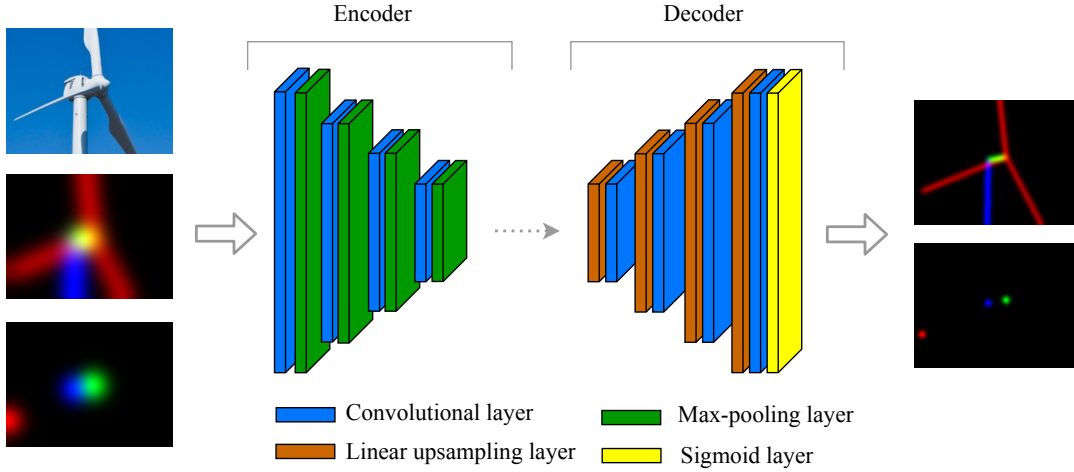


Fig. 2. Overview of the CNN network showing layer architecture, inputs and outputs

convolutional-deconvolutional architecture. The input is successively convolved and downsampled using Max-Pooling up to the bottleneck of the network. Next, the data is convolved and upsampled up to its original dimensions. This type of architecture is beneficial in that visual information from across the spatial range of the input is brought together in the deepest part of the network to provide feature rich information to the output of the deconvolutional part of the architecture. An visual overview of the network can be seen in Figure 2.

The role of the network is to take an image of a wind turbine as input and produce an equivalent image showing the inferred projection of the turbine skeleton model detailed in Section III-A. For the line model \mathcal{L}^c , this will be an image \mathbf{I}^L containing a line running up the tower, a line connecting the tower with the blades, and a line running along the centre of each of the blades. For the point model \mathcal{P}^c , this will produce an image \mathbf{I}^P with a point at the bottom of the tower, a point at the top of the tower, a point at the centre of the blades, and points at the tips of the blades. As the lines and points in the two models correspond to different parts of the wind turbine, we separate the corresponding outputs into different classes. For \mathbf{I}^L , the tower is one class, the connecting line another class, and the lines running along the blades a third class. We don't separate the individual blades into different classes due to their rotational symmetry. For \mathbf{I}^P , the classes are the tower base, the tower top, the blade centre and the blade tips. Again, all three blade tips are the same class.

For a typical convolutional-deconvolutional network, the input would consist of just the RGB image of the turbine. However, in this work, associated with each image is an estimation of the camera pose obtained from the GPS/IMUs on the drone. We use this information to act as a prior on the line and point locations, making prediction easier for the network. To do this, we construct the skeleton model, and project it through the pose estimate and camera intrinsics to find locations of each part on the input image. As the error in the pose estimates is not excessive, these projections will lie close to their true locations. To ease the work of the network, we apply Gaussian smoothing to the projections

and then append these channels onto the input image and feed it into the network. This means that the network input is a ten channel image. Three channels for the RGB image, three channels for the line model priors and four channels for the point model priors. Examples of the input and outputs can be seen in Figure 2.

To train the network, we obtained a set of ~ 1000 images of wind turbines from the internet. Each of the images in the data set was manually labelled by placing landmarks at the base of the tower, the top of the tower, the centre of the blades and the tips of the blades. These 2D locations correspond to the 3D locations of the points in \mathcal{P}^c . We then use these landmarks to generate the labels used during training. For the point-based label images, we set the 2D pixel location of the landmark in the correct image channel to 1, with the remaining pixels set to 0. We then apply a Gaussian smoothing kernel with $\sigma = 5$ to increase the spread of the landmark in the image. Each channel of the image is then renormalised to between 0 – 1. For the line-based label images, we draw lines on the images by connecting the landmarks in the same way the landmarks are connected in \mathcal{L}^c . Again, we apply a Gaussian kernel with $\sigma = 5$ to the images to increase the spread and then renormalise.

To generate the priors, we applied a set of random affine transformations to the image landmarks. This was done to replicate the amount of error we would expect to see in the GPS/IMU pose estimates during a live flight. After the transformations, we create the images in the same way that the label images are made, but apply a larger amount of Gaussian smoothing ($\sigma = 20$). For each training batch, we increased the variability in the training data through augmentation. This was done by applying random translations, rotations, scaling and cropping prior to the generation of the labels and priors. We trained the network using the Adam optimiser with a learning rate of 0.001. Training was stopped when the test loss started to diverge from the training loss after 2 days.

C. Pose Graph Optimisation

The problem of estimating the drone's 3D position and orientation is modelled as a pose graph. As the drone performs

an inspection, at regular intervals a new node or keyframe is added to the graph. This node contains an estimate of the absolute pose $\mathbf{T}_i = (\mathbf{R}_i, \mathbf{t}_i)$ obtained from the GPS/IMU, where \mathbf{R}_i is the orientation and \mathbf{t}_i is the 3D position. The aim is to optimise the graph using a set of constraints, such that the poses at the nodes converge to the drone's true location and orientation \mathbf{T}_i^* . The graph constraints are built using the pose estimates and the inferred projection of the 3D skeleton model produced by the CNN. Similar to most pose graph methods, the absolute pose estimate is not used during optimisation. Instead, it is used to compute the relative pose offset $\mathbf{T}_{i,i-1}$ between the current keyframe \mathbf{T}_i and the previous keyframe \mathbf{T}_{i-1} , i.e.

$$\mathbf{T}_{i,i-1} = \begin{bmatrix} \mathbf{t}_{i,i-1} \\ \mathbf{q}_{i,i-1} \end{bmatrix} = \begin{bmatrix} \mathbf{R}_i^T (\mathbf{t}_{i-1} - \mathbf{t}_i) \\ \mathbf{q}_i^{-1} * \mathbf{q}_{i-1} \end{bmatrix}, \quad (1)$$

where \mathbf{q}_i is the quaternion representation of the rotation \mathbf{R}_i . This is beneficial as over a short period of time there is less scope for error to accumulate in the GPS/IMU measurements. The image measurements \mathbf{I}_i^L and \mathbf{I}_i^P , are created using the CNN described in Section III-B, with the priors generated by projecting the skeleton model through the camera using \mathbf{T}_i and the camera intrinsics matrix \mathbf{K} .

To optimise the graph, we define a cost function that computes the residual error between the expected measurements – given the current state $\hat{\mathbf{T}}_i$ of the optimiser – and the sensor measurements described above

$$E = \sum_i e_{i,i-1}^T + \sum_i e_i^I. \quad (2)$$

The optimal set of poses are those which minimise this function. This function is optimised each time a new pose is added to the graph, and is initialised with the results from the previous optimisation. The first term $e_{i,i-1}^T$ compares the relative pose of the current estimates $\hat{\mathbf{T}}_{i,i-1}$ with those obtained from the GPS and IMU, i.e.

$$e_{i,i-1}^T = \mathbf{C} \begin{bmatrix} \hat{\mathbf{t}}_{i,i-1} - \mathbf{t}_{i,i-1} \\ 2 \times \text{Vec}(\hat{\mathbf{q}}_{i,i-1} * \mathbf{q}_{i,i-1}^{-1}) \end{bmatrix}, \quad (3)$$

where ‘Vec’ corresponds to the vector part of the quaternion rotation, and \mathbf{C} is a diagonal matrix weighting the different elements of the cost. Due to limitations in our flight software, the covariances of the relative pose measurements are unavailable. Instead we set the values in \mathbf{C} manually prior to the optimisation, based on the expected error in the IMU/GPS measurements, weighting each term in the translation error by β^t and those in the rotation error by β^q . The output $e_{i,i-1}^T$ is then a 6D vector containing the residuals in rotation and translation.

The cost function for the image measurements is based on point-to-point correspondences which are established differently depending on the types of image measurements. Figure 3 provides an illustration. To establish correspondences we adopt an active search approach as follows. For each of the point-based measurements \mathbf{I}_i^P , we project the points $\mathbf{p}_n^c \in \mathcal{P}^c$, using the current estimated pose of the camera

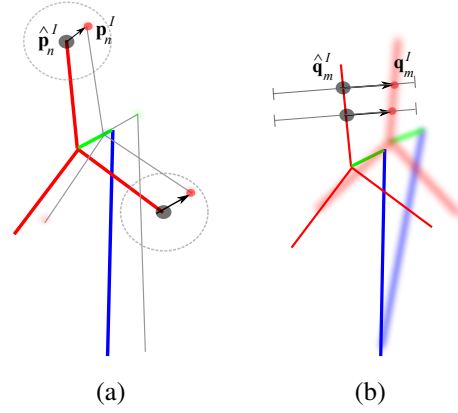


Fig. 3. Correspondence matching. (a) Point-based correspondences are found by searching within a circular window. (b) Line-based correspondences are found using a perpendicular line search. Model projections inferred by the CNN are shown as blurred points and lines and the model projection using the current pose estimate $\hat{\mathbf{T}}_i$ is shown in bold colours.

$\hat{\mathbf{T}}_i$ and the camera intrinsic matrix \mathbf{K} to find their location on the image plane $\hat{\mathbf{p}}_n^I$. Given this location, we extend a circular window of radius r^P and search the pixels within the window to find the one with the largest value which is selected as the correspondence \mathbf{p}_n^I , as shown in Fig. 3a. If there are no pixels with value above a threshold λ^P then no correspondence is established. This process is repeated for all the points in the turbine model and across all the point image measurements.

For the line-based measurements \mathbf{I}_i^L , the process is more complicated. First, we extract a set of points $\mathcal{Q}^c = \{\mathbf{q}_1^c, \dots, \mathbf{q}_m^c\}$ from the lines in \mathcal{L}^c . This is done by subdividing the lines at regular intervals. The number of subdivisions for each type of line is different due to their differing lengths. The tower is subdivided into s^t points, the hub into s^h points and the blades into s^b points. Next, we project each of the subdivided points \mathbf{q}_m^c into the image using $\hat{\mathbf{T}}_i$ and \mathbf{K} to get their 2D locations $\hat{\mathbf{q}}_m^I$. Instead of the circular search area described for the point-based correspondences, we instead do a perpendicular line search from the projected points as shown in Fig. 3b. The reason for this is that often the projected line, and the corresponding line in \mathbf{I}_i^L will be near parallel to one-another. We therefore want the correspondence to be the closest point perpendicular to the projected line.

To perform the perpendicular line search, for each line in \mathcal{L}^c , we project the two end points into the image space and find the 2D line connecting them. The 2D direction perpendicular to this line is the direction used during the line search. For each of the subdivided points $\hat{\mathbf{q}}_m^I$ we then sample the pixel values perpendicular to the line over a length of a^L with k^L sample locations. The pixel with the highest value is chosen as the correspondence \mathbf{q}_m^I . Again, if there are no values above a threshold λ^L , no correspondence is established. After finding the set of correspondences for each frame as described above, we can define the image cost function as

$$e_i^I = \beta^p \sum_n^{N_i} |\hat{\mathbf{p}}_n^I - \mathbf{p}_n^I| + \beta^q \sum_m^{M_i} |\hat{\mathbf{q}}_m^I - \mathbf{q}_m^I|, \quad (4)$$

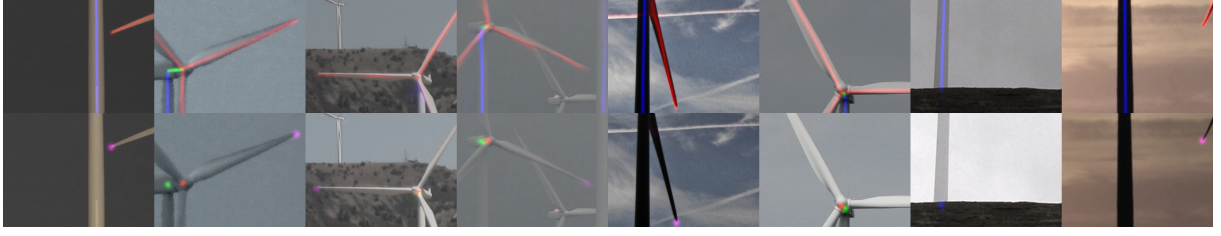


Fig. 4. Example output from the CNN. Top) Lines output \mathbf{I}^L . Bottom) Points output \mathbf{I}^P

where N_i represents to the number of point correspondences in the frame i and M_i represents to the number of line correspondences in frame i . The values β^p and β^q are used to weight the different types of correspondences.

With the cost function fully defined we are able to optimise the pose graph. This is done using the Gauss-Newton algorithm which works by linearising the problem around the current best guess solution, finding the minimum and repeating until convergence. As we are provided with initial pose estimates from the GPS / IMU, we found this method appropriate for our problem.

IV. EXPERIMENTS

As we do not have access to ground truth pose estimates for the inspection flights, we are unable to perform a quantitative evaluation of the overall performance of the method. However, we are able to evaluate the different sections in isolation. In Section IV-A, we evaluate the performance of the CNN, and show the importance of incorporating prior information into the network input. In Section IV-B we show the performance of the pose estimation part of the system using synthetic data. Finally, in Section IV-C we give a qualitative evaluation of the system using real-world inspection data.

A. Network Evaluation

Figure 4 shows example outputs from the CNN for 8 partial views of wind turbines, with the line and point estimates shown in the top and bottom rows, respectively. The colours indicated the different line and point classes. Note that even with a very limited view of the turbine, the network is able to accurately predict the projection of the lines and points of the skeleton model. We also evaluated the impact of including prior information about the turbine lines and points with the input on the CNN performance. We trained two versions of our architecture, one including prior information, and the other without prior information. The networks were identical apart from the shape of the filters in the first layer which enable the inclusion of the extra channels. The networks were trained using identical data sets and for exactly the same number of epochs. To evaluate the performance, we applied the networks to a set of test data and computed the pixel-wise mean squared error between the network predictions and the ground truths. The mean squared error for the network with priors was 0.001, and the mean squared error for the network without priors was 0.0012. This shows that performance is improved by

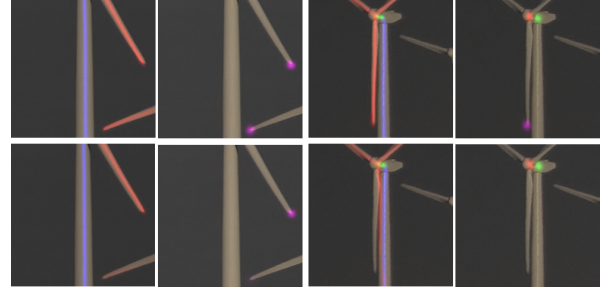


Fig. 5. Example line outputs from the evaluation of the effect of using prior in the CNN. Top row) Network with prior information, Bottom row) Network without prior information

including prior information. This is backed up in Fig. 5 where we compare the two network outputs. We can see that when the prior is included, the performance of the CNN is much more consistent, particularly in predicting the blade lines.

B. Simulated Evaluation

To provide an investigatory evaluation of the performance of the pose estimation part of the system, we designed an experiment using synthetic data. We first extracted the GPS/IMU poses from a set of actual inspection flights to give some example flight paths to use as ground truths. We next added progressive Gaussian noise to the ground truths on both the translations and rotations to provide us with an example of the sort of errors we would expect to accumulate over the course of an inspection. For the translations, we sampled a random 3D offset from a zero mean Gaussian distribution and added it to each of the nodes. For the rotation, we sampled a random angle from a zero mean Gaussian distribution, as well as a randomised normalised vector and applied this as an axis-angle rotation to each of the poses. This gave us the simulated IMU/GPS poses which at the start of the flight are close to the ground truth with the error getting progressively worse throughout the inspection. An example can be seen in Figure 7. These noisy poses are used as the relative pose measurements \mathbf{T} in the optimiser. To generate the image measurements, we directly simulate the output of the network by projecting the lines \mathcal{L}^c and points \mathcal{P}^c representations through the set of ground truth poses at each node of the graph. Although the simulated image measurements are error free – unlike the typical output of the network – these are suitable for evaluating the performance of the pose optimisation.

The measure that we are interested in evaluating is the robustness of the optimiser as increasing amounts of error are

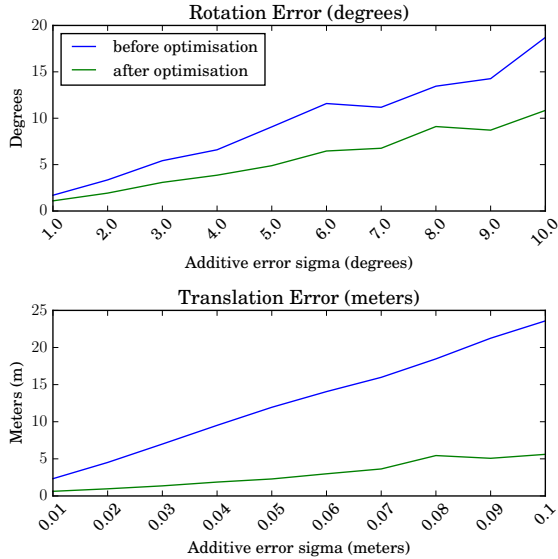


Fig. 6. Synthetic experiment results. Top) Pre and post optimisation rotation error. Bottom) Pre and post optimisation translation error.

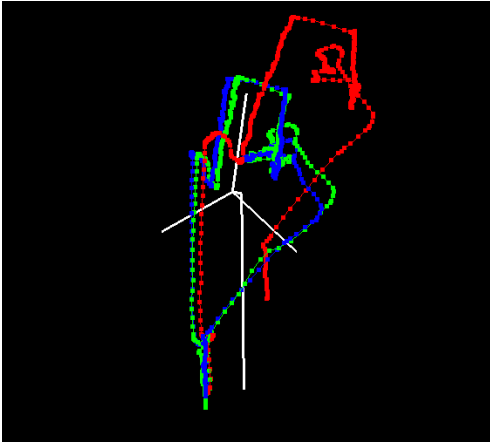


Fig. 7. Output from synthetic experiments. Ground truth path is in blue, input path is red, and the path after optimisation is in green.

added to the relative pose measurements. Knowing this gives us an understanding of how accurate the initial GPS/IMU pose measurements need to be to be able to reasonably correct the localisation. To evaluate this, we generated a series of synthetic flights as described above with translations errors ranging from 0.01m - 0.1m, and rotation errors from 1-10 degrees. We applied our method to these datasets and recorded the average translation error and rotation error for each flight. The results can be seen in Figure 6 and an example output can be seen in Figure 7. From the plots we can see that the system is able to handle quite a large amount of error in the pose measurements, especially for the translation. These preliminary results about the effectiveness of the method are encouraging, although more detailed analysis is needed to fully explore the failure states.

C. Visual Evaluation on Real Data

Our final experiment aims to provide a qualitative evaluation of the method applied to a real inspection flight.

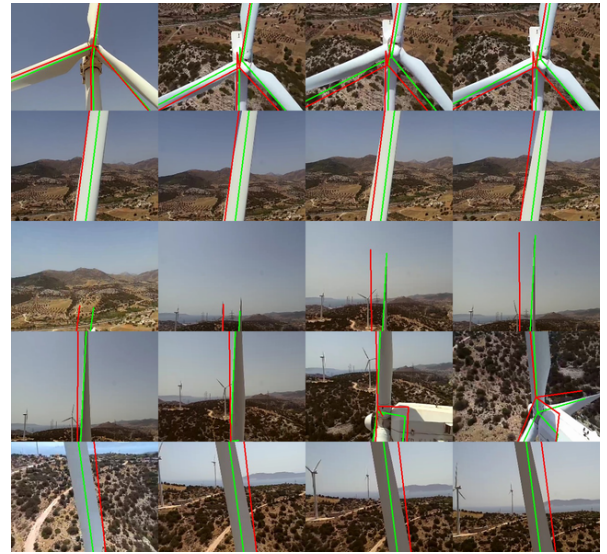


Fig. 8. Examples of the method being applied to real inspection data. The turbine model has been projected into the image using the pose prior to optimisation (red) after optimisation (green).

Although our method was applied after the flight had taken place, it was done in a way that exactly mimics how it would work if it were running during the flight. Examples of the optimiser output can be seen in Figure 8. From the images, we can see that the described method noticeably improves the localisation.

V. CONCLUSIONS

We have presented a novel method for integrating image-based measurements into drone localisation for the automated inspection of wind turbines. We have described a novel CNN-based system of producing a simplified representation of the wind turbine that allows for easy matching with our wind turbine model. We have also detailed how this representation is incorporated into a pose graph optimisation system. We evaluated the different sections of our work separately. We showed that the inclusion of prior information into the network, improved prediction performance by comparing it to a identical network without prior information. We also evaluated the performance of the pose estimation using synthetic data. Finally we gave a qualitative evaluation of the complete system when applied to an inspection flight.

There are a number of different avenues for future work. First, we would like to properly integrate the GPS/IMU measurements into the system, allowing us to obtain the covariances from these sensors which should improve the optimisation. We also intend to incorporate additional sensors such as LiDAR. Finally, we intend to would look at simultaneously estimating the parameters of the turbine models as part of the optimisation.

ACKNOWLEDGEMENTS

The authors acknowledge the support of Innovate UK (project number 104067). The work described herein is the subject of UK patent applications GB1815864.2 and GB1902475.1

REFERENCES

- [1] J. Ribrant and L. M. Bertling, "Survey of failures in wind power systems with focus on Swedish wind power plants during 1997-2005," *IEEE Transactions on Energy Conversion*, vol. 22, no. 1, pp. 167–173, mar 2007.
- [2] F. P. García Márquez, A. M. Tobias, J. M. Pinar Pérez, and M. Pappalías, "Condition monitoring of wind turbines: Techniques and methods," *Renewable Energy*, vol. 46, pp. 169–178, oct 2012.
- [3] L. Wang and Z. Zhang, "Automatic Detection of Wind Turbine Blade Surface Cracks Based on UAV-Taken Images," *IEEE Transactions on Industrial Electronics*, vol. 64, no. 9, pp. 7293–7309, sep 2017.
- [4] S. Leutenegger, S. Lynen, M. Bosse, R. Siegwart, and P. Furgale, "Keyframe-based visual-inertial odometry using nonlinear optimization," *International Journal of Robotics Research*, vol. 34, no. 3, pp. 314–334, mar 2015.
- [5] A. Kendall, M. Grimes, and R. Cipolla, "Posenet: A convolutional network for real-time 6-dof camera relocalization," in *Computer Vision (ICCV), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2938–2946.
- [6] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.
- [7] A. Kendall and R. Cipolla, "Modelling uncertainty in deep learning for camera relocalization," in *Robotics and Automation (ICRA), 2016 IEEE International Conference on*. IEEE, 2016, pp. 4762–4769.
- [8] —, "Geometric loss functions for camera pose regression with deep learning," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 3, 2017, p. 8.
- [9] R. Clark, S. Wang, A. Markham, N. Trigoni, and H. Wen, "Vidloc: A deep spatio-temporal model for 6-dof video-clip relocalization," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 1, no. 2, 2017, p. 3.
- [10] S. Wang, R. Clark, H. Wen, and N. Trigoni, "DeepVO: Towards end-to-end visual odometry with deep Recurrent Convolutional Neural Networks," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, May 2017, pp. 2043–2050.
- [11] G. Pavlakos, X. Zhou, A. Chan, K. G. Derpanis, and K. Daniilidis, "6-dof object pose from semantic keypoints," in *Robotics and Automation (ICRA), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2011–2018.
- [12] A. Newell, K. Yang, and J. Deng, "Stacked hourglass networks for human pose estimation," in *European Conference on Computer Vision*. Springer, 2016, pp. 483–499.
- [13] O. Moolan-Feroze and A. Calway, "Predicting Out-of-View Feature Points for Model-Based Camera Pose Estimation," *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 82–88, oct 2018.
- [14] M. Stokkeland, K. Klausen, and T. A. Johansen, "Autonomous visual navigation of Unmanned Aerial Vehicle for wind turbine inspection," in *2015 International Conference on Unmanned Aircraft Systems, ICUAS 2015*. IEEE, jun 2015, pp. 998–1007.
- [15] B. E. Schäfer, D. Picchi, T. Engelhardt, and D. Abel, "Multicopter unmanned aerial vehicle for automated inspection of wind turbines," in *24th Mediterranean Conference on Control and Automation, MED 2016*. IEEE, jun 2016, pp. 244–249.